Supplementary Material: Quantifying Tumor Heterogeneity in Whole-Genome and Whole-Exome Sequencing Data

Layla Oesper, Gryte Satas and Benjamin J. Raphael

Contents

1 Proofs Omitted from the Main Paper

Corollary 2.1. *Suppose* $\mathbf{C} \in \mathcal{C}_{m,n,k}$. *If there exists an* $i \in \{1, \ldots, m-1\}$ *such that for all* $t \in \{2, \ldots, n\}, c_{i,t} \geq c_{i+1,t}$ *and there exists a* $t \in \{2, \ldots, n\}$ *such that* $c_{i,t} > c_{i+1,t}$, *then* $\Phi(\mathbf{C}) = \emptyset$.

Proof. Now, we will proceed by contradiction. Assume the above conditions hold and $\Phi(\mathbf{C}) \neq \emptyset$. Then there exists some μ such that $(\mathbf{C}\mu)_i \leq (\mathbf{C}\mu)_{i+1}$. Since, $\mu_k > 0$ for all $k \in \{2, \ldots, n\}$ this implies that $c_{i,s}\mu_s \geq c_{j,s}\mu_s$ for all $s \in \{2, \ldots, n\}$ and $c_{i,t}\mu_t > c_{j,t}\mu_t$ for some $t \in \{2, \ldots, n\}$. However, this implies $(\mathbf{C}\mu)_i > (\mathbf{C}\mu)_{i+1}$, a contradiction. Hence, $\Phi(\mathbf{C}) = \emptyset$. □

Theorem 2.1. *Let* $\mathbf{C} = [c_{i,j}]$ *be an interval count matrix.* $\mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu)$ *is a convex function of* $c_{i,j}$.

In order to prove Theorem 2.1 we first need to build up some notation. We will then prove Theorem 2.1 in the situation where $n = 2$, which is then easily generalizable to $n > 2$. We start with the following real valued function defined when $n = 2$. Given a read depth vector $\mathbf{r} = (r_1, \ldots, r_{m+1})$, and a pair $(\mathbf{C}, \mu) \in \Omega_{m,2,k}$ we define $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu} : [0, \infty) \longrightarrow \mathbb{R}$ such that $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x) = \mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu}|\mathbf{X} = [\mathbf{C}; 2, x])$. Here $\widehat{\mathbf{X}\mu} = \frac{\mathbf{X}\mu}{|\mathbf{X}\mu|_1}$ is just the normalized version of $\mathbf{X}\mu$. To prove Theorem 2.1 for the case when $n = 2$, we just need to show that $\mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu}|\mathbf{X} = [\mathbf{C}; 2, x])$ is convex in x. Before we do so, we first need the following lemmas.

Lemma 1.1. $\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = -\sum_{j=1}^{m} r_j \log(p_j) + \alpha$ *is separable convex for* $\mathbf{p} \in \Delta_{m-1}$.

Proof. See the supplement of [7]. □

Lemma 1.2. *Let* $(\mathbf{C}, \mu) \in \Omega_{m,n,k}$ *and* $[a, b]$ *be a non-negative real valued interval. The set* $\mathcal{X} = \{\widehat{\mathbf{X}\mu}|\mathbf{X} = [\mathbf{C}; 2, x], x \in [a, b]\}$ *is a convex subset of* Δ_m.

Proof. We show that every element in $\mathcal{X}$ can be written as a convex combination of two particular elements of $\mathcal{X}$ and therefore defines a line in $\mathbf{R}^{m+1}$ (embedded in Δ_m) which is by definition a convex set. Let $\mathbf{A} \in \mathcal{S}$ such that $\mathbf{A} = [\mathbf{C}; 2, a]$ and let $\mathbf{B} \in \mathcal{X}$ such that $\mathbf{B} = [\mathbf{C}; 2, b]$. Notice that for any $x \in [a, b]$ there exists some λ where $0 \leq \lambda \leq 1$ such that $x = \lambda a + (1 - \lambda)b$. Therefore, any $\mathbf{X} \in \mathcal{X}$ can be written as $\mathbf{X} = \lambda\mathbf{A} + (1 - \lambda)\mathbf{B}$ for some λ where $0 \leq \lambda \leq 1$. We note the following two observations which can easily be verified for any $\mathbf{X} \in \mathcal{X}$ and corresponding λ.

(1) $|\mathbf{X}\mu|_1 = |(\lambda\mathbf{A} + (1 - \lambda)\mathbf{B})\mu|_1 = |\lambda\mathbf{A}\mu + (1 - \lambda)\mathbf{B}\mu|_1 = \lambda|\mathbf{A}\mu|_1 + (1 - \lambda)|\mathbf{B}\mu|_1$.
(2) $(\mathbf{X}\mu)_i = \lambda(\mathbf{A}\mu)_i + (1 - \lambda)(\mathbf{B}\mu)_i$ for all $i \in \{1, \ldots, m+1\}$.

We now show that for any $\mathbf{X} \in \mathcal{X}$ there exists some $\alpha = (\alpha_1, \alpha_2) \in \Delta_1$ such that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$. Set $\alpha_1 = \frac{\lambda|\mathbf{A}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}$ and $\alpha_2 = \frac{(1-\lambda)|\mathbf{B}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}$. By definition $\alpha_1 + \alpha_2 = 1$ and $\alpha_1, \alpha_2 \geq 0$, so $\alpha \in \Delta_1$. We now show that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$. For each $i \in \{1, \ldots, m+1\}$ we compute the i^{th} entry:

$$\begin{aligned}
(\alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu})_i &= \alpha_1\frac{(\mathbf{A}\mu)_i}{|\mathbf{A}\mu|_1} + \alpha_2\frac{(\mathbf{B}\mu)_i}{|\mathbf{B}\mu|_1} \\
&= \frac{\lambda|\mathbf{A}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}\frac{(\mathbf{A}\mu)_i}{|\mathbf{A}\mu|_1} + \\
&\quad \frac{(1-\lambda)|\mathbf{B}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}\frac{(\mathbf{B}\mu)_i}{|\mathbf{B}\mu|_1} \\
&= \frac{\lambda(\mathbf{A}\mu)_i + (1-\lambda)(\mathbf{B}\mu)_i}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1} \\
&= \frac{(\mathbf{X}\mu)_i}{|\mathbf{X}\mu|_1} \text{ (Using both of the above observations.)} \\
&= (\widehat{\mathbf{X}\mu})_i
\end{aligned}$$

Hence, we see that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$, and is therefore any $\widehat{\mathbf{X}\mu} \in \mathcal{X}$ is a convex combination of $\widehat{\mathbf{A}\mu}, \widehat{\mathbf{B}\mu} \in \mathcal{X} \subseteq \Delta_m$. And therefore $\mathcal{X}$ must be a convex subset of Δ_m. □

We now can prove Theorem 2.1.

Proof. We start by considering the case where $n = 2$. Lemma 1.2 tells us that for a fixed $(\mathbf{C}, \mu) \in \Omega_{m,n,k}$ and closed positive real valued interval $[a, b]$ the set $\mathcal{X} = \{\widehat{\mathbf{X}\mu} | \mathbf{X} = [\mathbf{C}; 2, x], x \in [a, b]\}$ is a convex subset of Δ_m. Notice that $[0, \infty) = \cup_{i=1}^{\infty}[0, i]$ is the union of a non-decreasing sequence of convex intervals. Let $\mathcal{X}_i = \{\widehat{\mathbf{X}\mu} | \mathbf{X} = [\mathbf{C}; 2, x], x \in [0, i]\}$ and $\mathcal{X} = \cup_{i=1}^{\infty}\mathcal{X}_i$. From the proof of Lemma 1.2, it is clear that $\mathcal{X}_i \subset \mathcal{X}_{i+1}$ for all $i \geq 1$. Hence, each $\mathcal{X}_i$ is a non-decreasing sequence of convex subsets of Δ_m and therefore $\mathcal{X}$ is a convex subset of Δ_m where $\mathcal{X} = \{\widehat{\mathbf{X}\mu} \mid \mathbf{X} = [\mathbf{C}; 2, x], x \in [0, \infty)\}$.

Since $\mathcal{X}$ is a convex subset of Δ_m we can apply the result from Lemma 1.1 to prove that $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is separable convex for $\mathbf{p} \in \mathcal{X}$. Since $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x) = \mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu} \mid \mathbf{X} = [\mathbf{C}; 2, x])$ there is a one-to-one correspondence between x and $\mathbf{p} \in \mathcal{X}$, we have shown that $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x)$ is convex in x.

We have therefore shown that $\mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu)$ is convex in $c_{i,j}$ when $n = 2$. The proof easily extends to the case when $n > 2$ by determining an appropriate pair $(\mathbf{X}', \mu')$ where $\mathbf{X}' = (\mathbf{x}_1, \mathbf{c}_j) \in \mathbf{R}^{m,2}$ and $\mu' = (1 - \mu_j, \mu_j)$ and $\widehat{\mathbf{X}'\mu'} = \widehat{\mathbf{C}\mu}$ and the proof from the case of $n = 2$ can be directly applied. □

2 Using a Graph to Enumerate $\mathcal{S}_{m,n,k}$

In this section we provide further details and pseudocode on our algorithm for using a graph to enumerate $\mathcal{S}_{m,n,k}$.

Algorithm 1: Enumerate $\mathcal{S}_{m,n,k}$ using modified depth-first search on $G_{3,k}$. $\mu Set\,(v, w)$ is the set of values for μ for which $v\mu \leq w\mu$.

```
Input: $G_{n,k}$, $m$
Output: The set $\mathcal{S}_{m,n,k}$
procedure Setup ($G_{n,k}$, $m$)
    $\mathcal{S} \leftarrow \emptyset$
    $\mathbf{C} \leftarrow n \times m$ matrix
    $\Phi \leftarrow \emptyset$
    $V, E \leftarrow G_{n,k}$
    for $v \in V$ do
        $\mathbf{C}[1:] \leftarrow v$
        $\mathcal{S} \leftarrow \mathcal{S} \cup$ Enumerate ($C$, $1$, $\Phi$, $m$, $G_{n,k}$)
    return $\mathcal{S}$
procedure Enumerate ($C$, $i$, $m$, $\Phi$, $m$, $G_{n,k}$)
    if $i$ = $m$ then
        return $\mathcal{S} \cup \mathbf{C}$
    $V, E \leftarrow G_{n,k}$
    $v \leftarrow \mathbf{C}[i:]$
    for $(v, w) \in E$ do
        $\Phi \leftarrow \Phi \cup \mu Set(v, w)$
        if $\Phi \neq \emptyset$ then
            $\mathbf{C}[i+1:] \leftarrow w$
            $\mathcal{S} \leftarrow \mathcal{S} \cup$ Enumerate ($C$, $i+1$, $\Phi$, $m$, $G_{n,k}$)
    return $\mathcal{S}$
```

The algorithm depends on being able to calculate and efficiently union the $\mu Set\,(v, w)$, i.e the set of values for μ for which $v\mu \leq w\mu$. In the case where $n = 3$, this set is defined by the single variable, $\frac{\mu_2}{\mu_3}$.

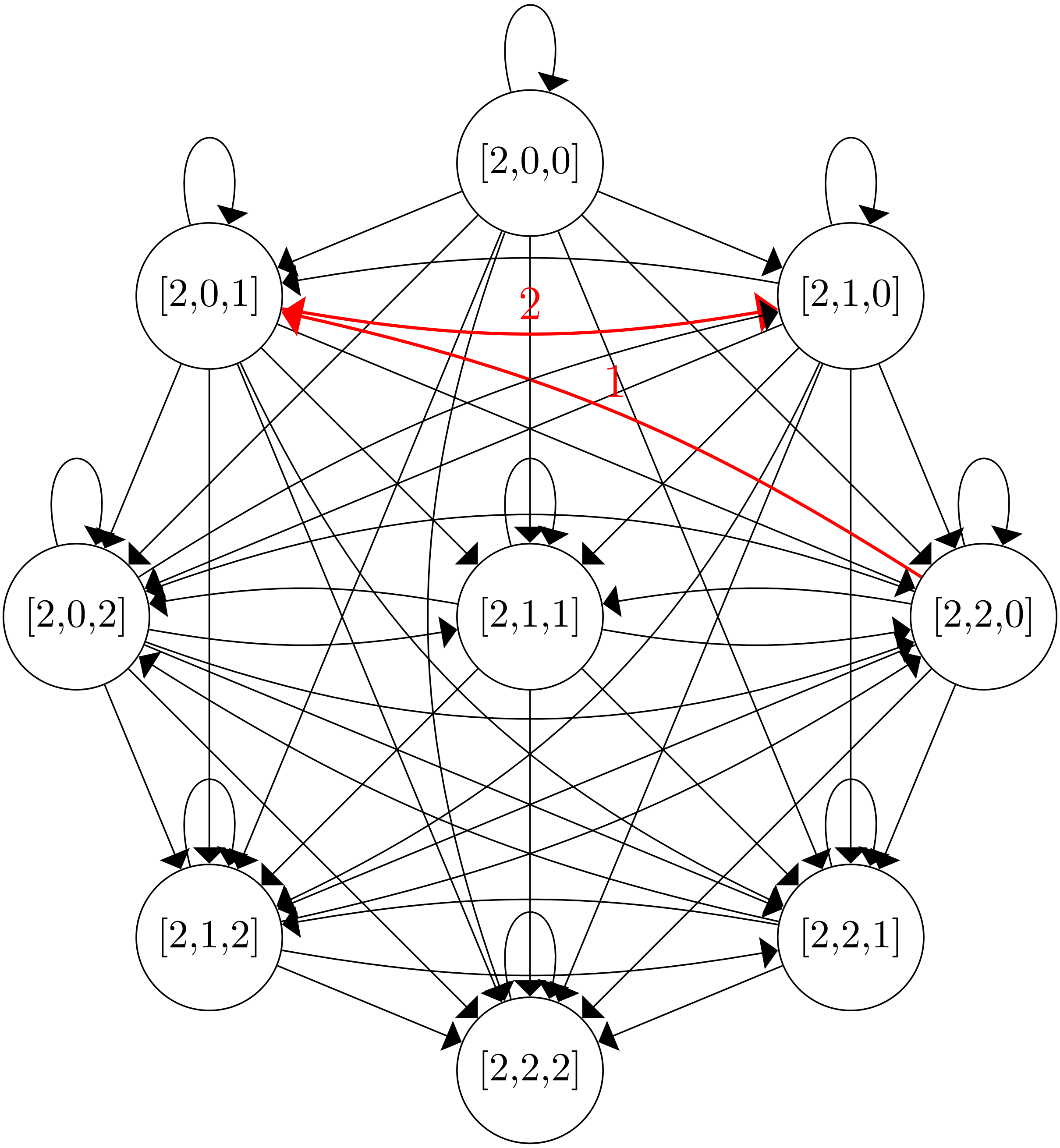


Figure S1: **Enumeration Graph for k=2**. Edges that cannot possibly result in valid matrices have been removed from the fully connected graph. However, a simple enumeration of the paths on this graph would still result in matrices which do not satisfy the compatible ordering condition (i.e. the set $\Phi(\mathbf{C})$ is empty). One example is the path highlighted in red. To account for this, the enumeration algorithm keeps track of Φ as paths are being enumerated and does not explore paths that cannot lead to valid matrices $\mathbf{C}$.

In particular, in the case where $v_2 > w_2$, the upper bound on $\frac{\mu_2}{\mu_3}$ is $\frac{w_3-v_3}{w_2-v_2}$. Likewise, in the case where $v_2 < w_2$, the lower bound on $\frac{\mu_2}{\mu_3}$ is $\frac{w_3-v_3}{w_2-v_2}$. The case where $v_2 = w_2$ doesn't restrict the values of $\frac{\mu_2}{\mu_3}$.

3 Interval Selection

In this section we discuss how interval selection is done during the first step of our two-step procedure for different values of n.

3.1 Mixtures of normal and one tumor subpopulation $(n = 2)$

For the first step in our two step procedure, we need a way to select a subset of high confidence intervals that will be used to infer $(\mathbf{C}^*, \mu^*)$ for just those intervals. Since we are modeling a sequencing experiment as a probabilistic model where reads are distributed according to a multinomial model, intervals with larger read depths are a natural candidate for selection. However, this may be confounded for intervals that are extremely amplified, thus resulting higher read counts, but where precise estimates of copy number are difficult to make. Therefore, we choose the intervals that have the longest length in the reference genome as a compromise between these competing interests. For a fixed integer d, we select up to the d longest intervals such that: (1) The number of tumor reads (t_j) and normal reads (n_j) aligning to interval I_j is non-zero; (2) The length of interval I_j is longer that $1Mb$; and (3) If T is the total number of tumor reads, N is the total number of normal reads and k is the provided max copy number parameter, then the following holds: $\frac{t_j/T}{n_j/N} < \frac{k+1}{2}$. This final constraint forces the observed copy number ratio to not be too high beyond the specified max copy number k. Additionally, if the set of selected intervals must represent $> 10\%$ of the total length of all provided intervals, otherwise the sample is determined to not be a good candidate for analysis using THetA. By default we set $d = 100$.

3.2 Mixtures of normal and two tumor subpopulations $(n = 3)$

When considering a tumor to be a mixture of multiple distinct tumor subpopulations ($n \geq 3$) we rely upon the results obtained from considering the tumor to only contain a single tumor population ($n = 2$) to find the set of intervals that allow us to best be able to measure events that have occurred in a subpopulation of tumor cells. In particular, we include intervals determined by the $n = 2$ analysis to have normal copy ($c^*_{j2} = 2$) as well as intervals determine to contain copy number aberrations ($c^*_{j2} \neq 2$). We also limit the copy number aberrations used to have either been predicted to be a deletion or an amplification of a single copy, as these intervals have the most reliable signal for predicting multiple tumor populations. For a fixed integer d (we use 20 by default) the interval selection process goes as follows:

- Select the top $a = \lceil d \times 0.75 \rceil$ longest intervals such that: (1) The length of the interval I_j is longer than $5Mb$, (2) $c^*_{j2} \neq 2$, and (3) $c^*_{j2} < 4$. If a such intervals do not exists, the genome is determined to not be a good sample for multiple tumor population analysis.
- Select the top $d - a$ longest intervals such that: (1) The length of the interval I_j is longer than $5Mb$, and (2) $c^*_{j2} = 2$. If $d - a$ such intervals do not exists, the genome is determined to not be a good sample for multiple tumor population analysis.

4 Determining Additional Copy Numbers: Multiple Rows

Individually estimating optimal copy numbers for low confidence intervals is not guaranteed to find optimal solution as if all intervals were jointly estimated. In order to test how well the procedure does in practice, we ran the two-step algorithm and inferred copy numbers for all intervals on three less fragmented (< 200 interval) whole genome and exome samples, as well as a single chromosome of a more fragmented sample (See Table 4). We then fixed the values of the high confidence intervals used in Step 1, and the estimated μ value, and through brute-force enumeration, found the true optimal value for the low confidence intervals. We find that for the less fragmented whole genome and exome samples, the step two procedure correctly inferred the optimal copy number for all intervals. On the single chromosome sample, the procedure was correct for all but one interval.

ID	Data Type	# Intervals (Total)	# Intervals (Step 2)	# Intervals Incorrect
TCGA-06-0137	Exome	163	75	0
TCGA-AO-A0JF	WGS (low)	129	29	0
TCGA-BH-A0W5	WGS (low)	53	3	0
TCGA-56-1622 (Chrm 1)	WGS	159	92	1

Table S1: **Step Two Optimality.** Comparison of copy numbers inferred during step two of the two-step procedure to the optimal values.

5 Probabilistic Model of BAFs

In this section we describe in more detail our probabilistic model of BAFs. Let $\mathbf{s} = (s_1, s_2, \ldots, s_q)$ be a set of genomic coordinates for germline heterozygous SNPs in a patient, and let $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ be the observed BAFs across all $s \in \mathbf{S}$ in the normal sample and $\mathbf{w} = (w_1, w_2, \ldots, w_q)$ be the BAFs for the corresponding tumor sample. We use a probabilistic model to describe $\mathbf{w}$.

Assuming that reads are generated uniformly at random across all DNA in a sample, we first calculate the expected deviation in BAF away from 0.5 for an interval I_j given a pair $(\mathbf{C}, \mu) \in \Omega$. In order to calculate this deviation, we need to know the number of copies of I_j for both parental chromosomes (and hence the number of copies of each allele for any $s_i \in I_j$. We do not need to know which copy number pertains to which allele, just the pair of integer values. We note that if we assume that if we make the simplifying assumption that no region is deleted, followed by a gain (and vice versa), we can exactly determine these values for regions of total copy $0, 1, 2$ and 3. Therefore, for the remainder of this section we assume that no entry in $\mathbf{C}$ is greater that 3. We define a function ϕ that given total copy number of an interval, returns the number of copies of the more common parental chromosome. That is $\phi(0) = 0$, $\phi(1) = 1$, $\phi(2) = 1$, and $\phi(3) = 2$. We now can define a value δ_j that gives the deviation away from 0.5 expected for interval I_j given a pair $(\mathbf{C}, \mu)$:

$$\delta_j = \frac{\sum_{k=1}^{n} \phi(c_{jk})\mu_k}{\sum_{k=1}^{n} c_{jk}\mu_k} - 0.5. \tag{1}$$

That is, δ_j is the fraction of total copies of interval I_j that contain the major (or more common) allele for any germline SNP located in I_j. For example, if interval I_j has not undergone any copy number events $c_{jk} = 2$ for all k then $\delta_j = 0$. Let $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$ be the expected deviation away from 0.5 for all intervals in $\mathbf{I}$. Note that if $\delta_j \neq 0$ we expect that the BAFs in interval I_j will be double banded, containing two clusters around $0.5 \pm \delta_j$.

We define a map $\pi : \{1, \ldots, q\} \longrightarrow \{1, \ldots, m\}$ where $I_{\pi(i)} \in \mathbf{I}$ is the genomic interval that contains SNP s_i. Let $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ where σ_j^2 is the observed variance around 0.5 for all heterozygous SNPs in interval I_j in the matched normal genome. That is $\sigma_j^2 = \frac{\sum_{i=1}^{q} \mathbb{1}(\pi(i),j)(v_i - 0.5)^2}{\sum_{i=1}^{q} \mathbb{1}(\pi(i),j)}$ where $\mathbb{1}$ is the identity function. Lastly, we define a sign function $sgn(x)$ such that $sgn(x) = 1$ if $x \geq 0$ and $sgn(x) = -1$ if $x < 0$. We now present a probabilistic model using a collection of gaussians for observed BAFs $\mathbf{w}$ given a pair $(\mathbf{C}, \mu) \in \Omega$ and observed BAFs in the matched normal $\mathbf{v}$ as a product of draws from different normal distributions.

$$P(\mathbf{w}|\mathbf{C}, \mu, \mathbf{v}) = P(\mathbf{w}|\delta, \sigma^2) = \prod_{i=1}^{q} P(w_i|\delta, \sigma^2) = \prod_{i=1}^{q} \mathcal{N}(w_i; 0.5 + sgn(w_i - 0.5)\delta_{\pi(i)}, \sigma_{\pi(i)}^2) \tag{2}$$

Given multiple pairs $(\mathbf{C}, \mu)$ with the same likelihood using only read depth, we may select the pair that maximizes the likelihood in Equation (2) to select the reconstruction most consistent with observed BAF data.

6 Simulations

6.1 Simulation Procedure

We create a simulated mixture of a specified number of tumor subpopulations along with normal admixture using real sequencing data from an AML tumor sample and matched normal sample (TCGA-AB-2965) from The Cancer Genome Atlas [1]. This sample was chosen due to its high purity (approximately 95% pure) and lack of copy number aberrations. We create tumor subpopulations similar to the glioblastoma genomes analyzed in the next sections by using up/down sampling to randomly spike in chromosome arm deletions and amplifications (we excluded the p-arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 from consideration). For each mixture we ensure that some aberrations are shared by different populations and that some are unique to the subpopulations. We then created a mixture by selecting reads uniformly at random from the original tumor genome and the created subpopulations to create a simulated mixtures. We then used the true matched normal sample as the normal sample in the simulation. Using up and down sampling we can create mixtures of different coverages.

We run our simulated data through the same pipeline as real data, including interval partitioning determine by using BIC-seq [11]. We note that BIC-seq recommends using a parameter setting of $\lambda = 2$ for low-coverage genomes and $\lambda = 4$ for higher coverage genomes. We adhere to these recommendations for these simulations.

6.2 Additional Simulation Results - Mixtures made with Normal Only

We note that in the main manuscript and this supplement, we include simulated data where a mixture was created by spiking in deletions and amplifications into a tumor sample which are then mixed with the original tumor sample and compared against the normal sample. As validation we also created similar mixtures by using the normal sample for all steps. We note that the data created by such a procedure will not include variation present in real data such as batch effects. We find that mixtures created using only the matched normal sample are segmented into many fewer intervals ($<$100) than when the mixture is created using the tumor sample (1000's intervals). As a result, we also find that THetA2 is able to perfectly reconstruct both μ and $\mathbf{C}$ for the mixtures created using only the normal sample (perhaps due to the fewer number of intervals). Therefore, we find that such simulations are valid for demonstrating that the implementation of THetA2 works as expected, but do not represent a realistic simulation given what we would expect to find in real sequencing data.

6.3 Additional Simulation Results - 7X Coverage

We also generated simulated data with 7X coverage. We find similar trends in 7X sequence coverage data as we see with 30X sequence coverage. Namely, we find good performance at estimating μ, the larger tumor population (Tum1) and increased performance at estimating the copy numbers in the smaller tumor population (Tum2) as it increases in size. We also see increased performance at estimating copy number aberrations in both tumor populations when only considering longer intervals (Supplemental Fig S2a). We comparing to 30X coverage simulations we see similar results, except with improved performance at estimating copy numbers for longer intervals (Supplemental Fig S2b).

6.4 Additional Simulation Results - Comparison of THetA to THetA2

We include here additional results from comparing THetA to THetA2 on simulated data. We use simulated mixtures of 3 subpopulations where the proportion of the sample in the larger tumor subpopulation

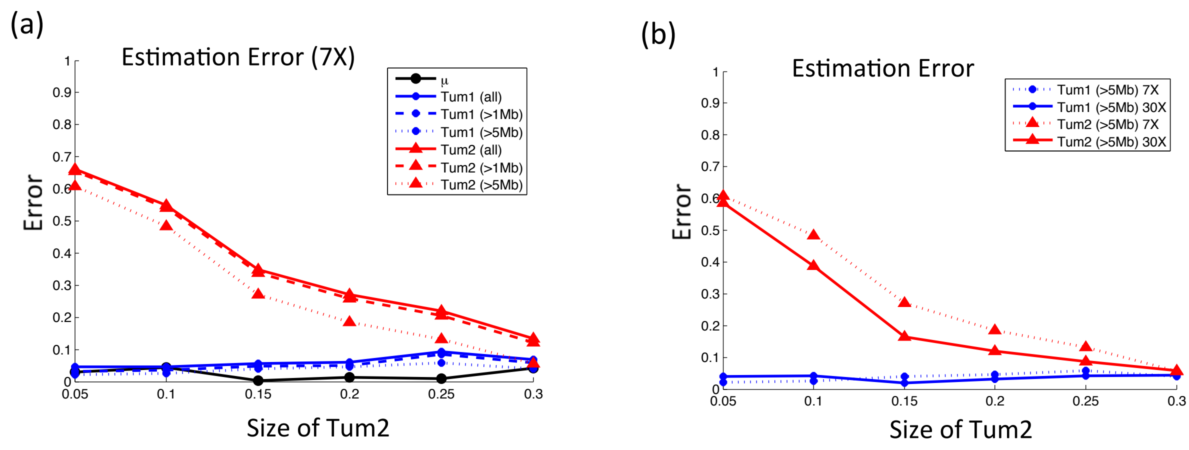


Figure S2: **Simulations with 7X Coverage and Comparison to 30X (a)** Estimation error for both μ and **C** for each tumor population (Tum1 and Tum2) as the size of Tum2 increases and the size of Tum1 is fixed at 0.5 for 7X coverage. Error in μ is euclidean distance from the true μ and error for each tumor population is the fraction of the genome for which the copy number is incorrectly inferred. We also report error rates for estimating copy numbers in both populations when we only restrict consideration to longer intervals. **(b)** Comparison of 30X to 7X coverage when considering intervals longer than 5Mb.

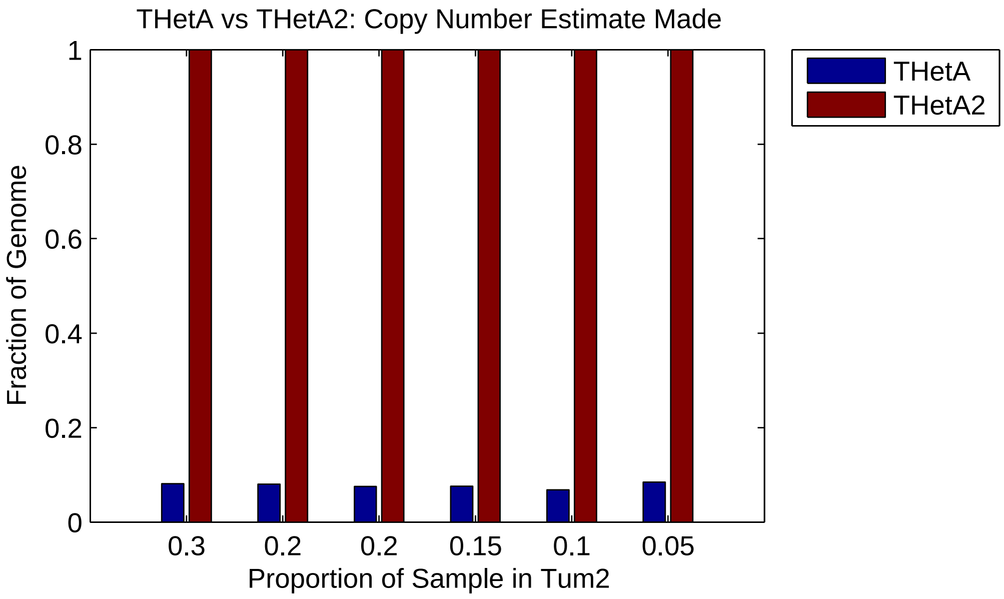


Figure S3: **THetA vs THetA2: Fraction of Genome Considered**. A bar plot showing the fraction of the genome for which copy number estimates are made for both THetA and THetA2.

is fixed at 0.5 and the proportion of the sample in the smaller subpopulation varies from 0.05 to 0.3. Figure S3 shows a comparison between what fraction of the genome THetA and THetA2 make copy number estimates. The two-step procedure allows THetA2 to consider all of the genome while THetA only considers less than 10% of the genome.

On these same simulations we also compare the accuracy of estimating both μ and **C**. When no copy number prediction is made for a region, we consider this to be an incorrect prediction.

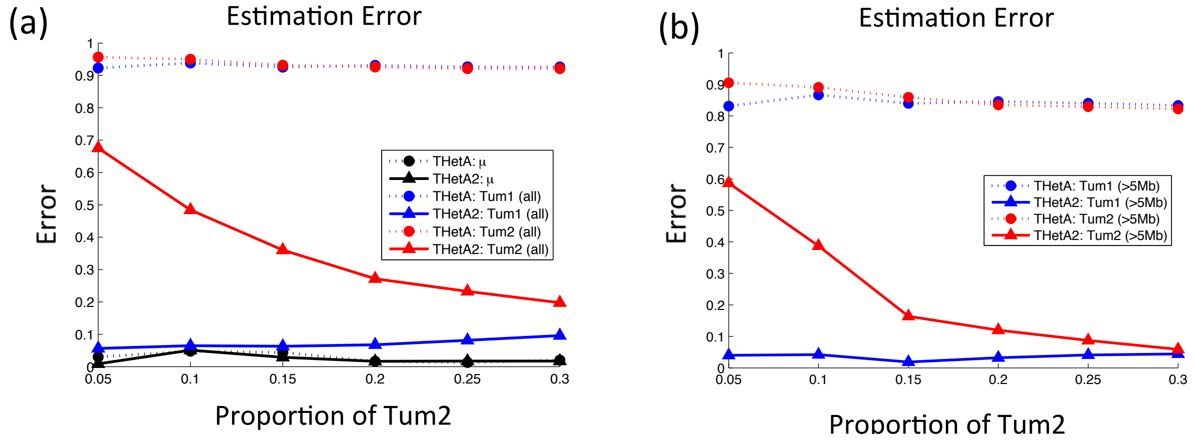


Figure S4: **Comparison of THetA vs THetA2**. (Left) Comparison of error at estimation μ (measured as euclidean distance from true) and **C** (measured as fraction of genome incorrectly estimated) between THetA and THetA2. (Right) Comparison of error at estimating **C** between THetA and THetA2 when only restricting consideration to intervals longer than $5Mb$.

6.5 Additional Simulation Results - Mixture of 4 subpopulations

We generated a mixture containing one normal and three tumor subpopulations. Subpopulation sizes were chosen to be sufficiently distinct from one another (20%, 30%, 40%). Whole arm deletions and amplifications were spiked into the mixture.

Due to the additional runtime requirements for 4 subpopulations, an alternate segmentation procedure was used. The simulated sample was divided into 50 kb intervals. We filtered out intervals which were likely to be noisy or lower quality: ones within the centromeres, ones that contained less than 2000 reads from the normal sample, and ones for which the ratio of tumor to normal reads was greater than 10% different from both of its neighbor intervals were filtered out, leaving 87.6% of the genome. For each chromosome, kernel density estimation of the distribution of tumor to normal read ratios was used to cluster intervals into larger intervals that we expect to contain the same copy number, then these large intervals were merged with intervals from other chromosomes which display similar tumor to normal read ratios.

Supplemental Fig. S5 shows the results of running THetA2 on these intervals. THetA2 was able to infer μ with 4.9% estimation error (using euclidian distance from the true μ. THetA2 was also able to infer the correct copy number values for 99.9%, 99.9%, and 99.6% of the intervals that were considered for the 3 tumor subpopulations respectively, which cover ~87.6% of the whole genome.

6.6 Additional Simulation Results - Underestimating Number of Subpopulations

We also investigated THetA2's behavior when the number of tumor subpopulations is incorrectly estimated. We considered six different mixtures of 3 subpopulations and evaluated the results returned by THetA2 when the number of subpopulations was fixed at two ($n = 2$). We find that THetA2 consistently underestimates tumor purity, but only by 0.027 on average (Supplemental Fig. S6). We also compared the values in the integer count matrix **C** returned by THetA2 to the true **C** values for the large and small subpopulations. We find that in this case, THetA2 was able to accurately estimate the copy number profile of the major subpopulation: on average 97.0% of the whole genome, and 96.6% of aberrant regions (regions which contain an amplification or deletion in at least one subpopulation). Thus, THetA2 may return useful information about a sample's purity and copy number profile, even if runtime constraints force THetA2 to underestimate the true number of subpopulations.

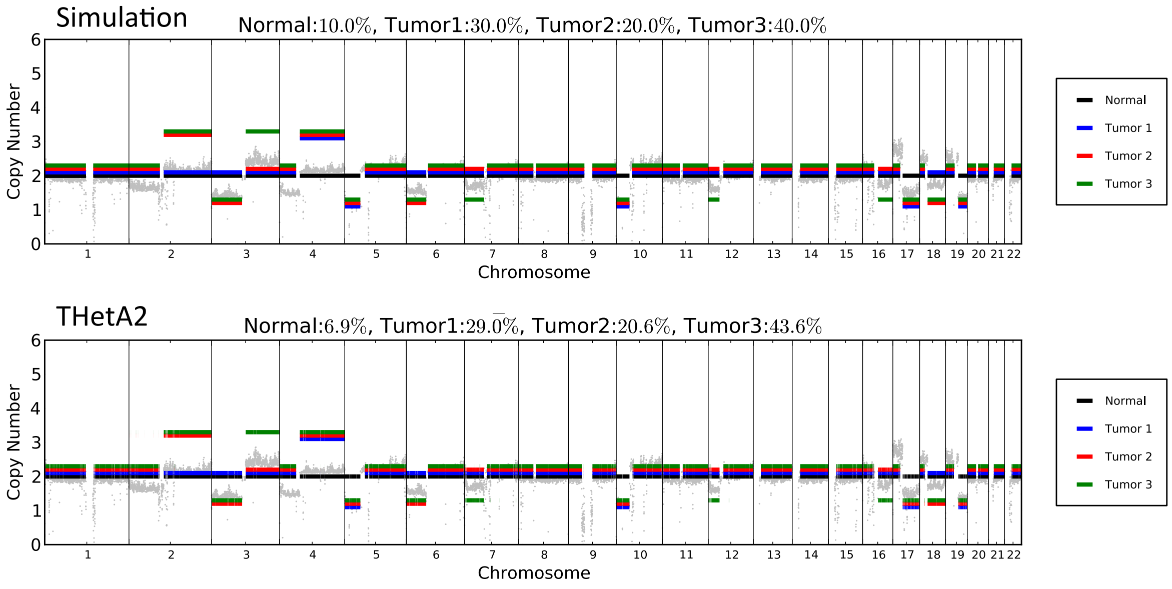


Figure S5: **Simulation with** 4 **subpopulations** The simulated mixture was created by spiking in chromosome arm deletions and amplifications to create three distinct tumor populations and mixing with a matched normal genome. Due to runtime concerns for $n = 4$, an alternative segmentation algorithm was used to obtain the intervals used. The figure shows the true mixture (above) and the solution obtained by THetA2 (below).

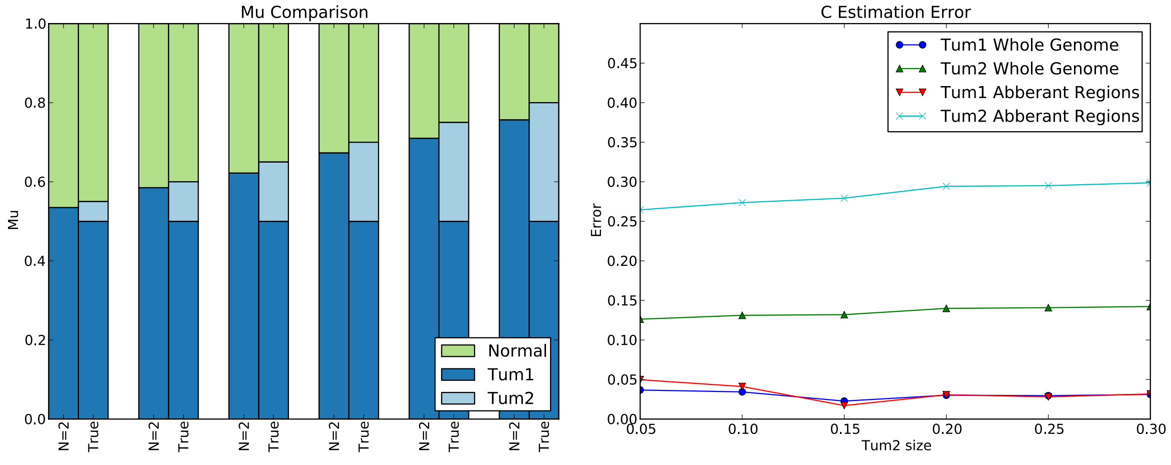


Figure S6: **THetA2 results when underestimating the number of subpopulations.** We ran THetA2 with the number of subpopulations fixed at two ($n = 2$) on six simulated 30X mixtures of 3 subpopulations. (a) For each mixture, the predicted μ is shown next to the true underlying μ. We find that the THetA2 tends to slightly underestimate the tumor purity when considering fewer subpopulations than exist in the true underlying mixture. (b) For each mixture, the copy number profile $\mathbf{C}$ predicted was compared to the true copy number profile for the large and small subpopulation. The fraction of the genome estimated incorrectly is shown, for both the whole genome and the aberrant regions (those that contain an amplification or deletion in at least one subpopulation). We find that when considering fewer subpopulations than exist in the true mixture, THetA2 copy number predictions tend to resemble those in the largest true subpopulation.

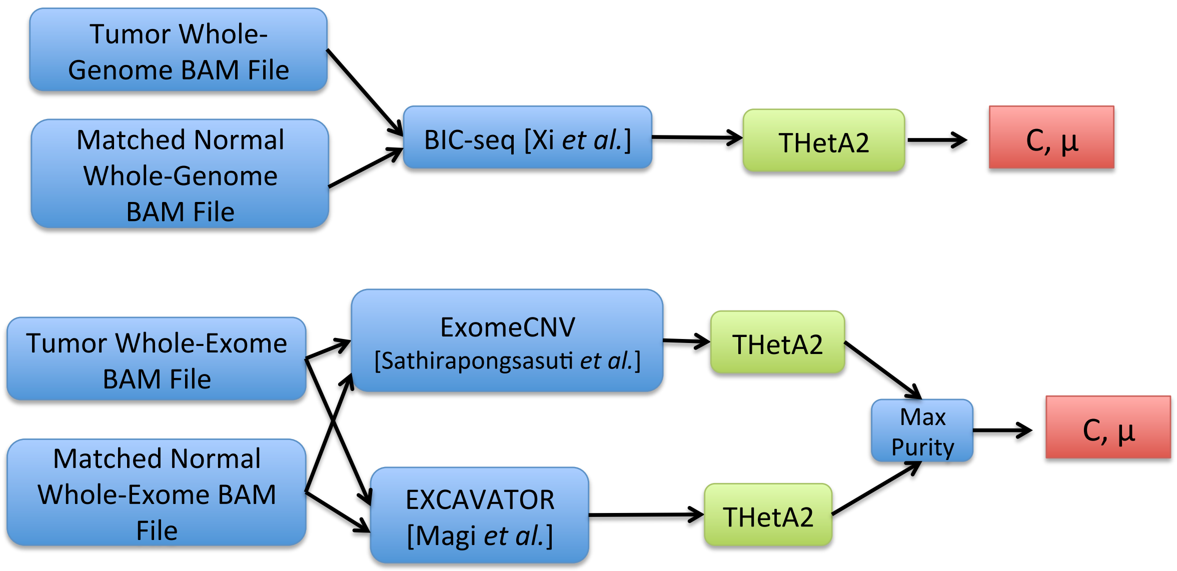

Tumor Whole-Genome BAM File
Matched Normal Whole-Genome BAM File
BIC-seq [Xi *et al.*]
THetA2
C, μ
Tumor Whole-Exome BAM File
Matched Normal Whole-Exome BAM File
ExomeCNV [Sathirapongsasuti *et al.*]
EXCAVATOR [Magi *et al.*]
THetA2
THetA2
Max Purity
C, μ

Figure S7: **Workflow for whole-genome and whole-exome datasets.**

7 Real Data Processing

7.1 Whole-Exome Data

BAM files for each sample were obtained from CGHub (`https://cghub.ucsc.edu/`) and only reads with a mapping quality ≥ 30 were used in our analysis. We determined exon positions $\mathbf{E}$ using the UCSC genome browser [5] and merging any overlapping exonic intervals. For each sample we used both ExomeCNV [8] and EXCAVATOR [6] run with default parameters to determine an interval partition $\mathbf{I}$. ExomeCNV directly provides a segmentation $\mathbf{I}$. Whereas, EXCAVATOR only provides regions that were predicted to contain non-normal copy, so $\mathbf{I}$ was determined to be the set of returned intervals and all genomic segments located between returned intervals. Read depth $\mathbf{r}$ over these intervals and the set of exons $\mathbf{E}$ was calculated as described in the methods for both the tumor and normal genomes. A diagram describing the workflow for whole-exome data is shown in Fig. S7.

7.2 Whole-Genome Data

BAM files for each sample were obtained from CGHub (`https://cghub.ucsc.edu/`) and concordant reads (as determined by using the GASV pre-processing utility [9]) with a mapping quality $\geq$ 30 were used in our analysis. For each sample we used both BIC-seq [11] run with default parameters ($\lambda = 2$) to determine an interval partition $\mathbf{I}$. A diagram describing the workflow for whole-genome data is shown in Fig. S7.

7.3 Virtual SNP Array

See our previous publication [7] for details of how a virtual SNP array is created. When comparing results obtained from THetA2 to data from a virtual SNP array, we calculate an observed mean BAF. This value is calculated for each interval in an interval partition of the genome obtained from either the BIC-Seq [11], ExomeCNV [8] or EXCAVATOR [6]. We only report values for intervals that are longer than 2Mb and contain at least 10 heterozygous SNPs in the matched normal sample. We calculate the standard deviation in the observed B-allele frequencies (BAFs) for all germline SNPs occurring in the specified interval in both the tumor (σ_t) and matched normal (σ_n) samples. If the $\sigma_t < 1.5\sigma_n$, then we report the mean as 0.5, as would be expected in a non-rearranged interval. However, if $\sigma_t >= 1.5\sigma_n$, then we report two mean values - the mean of all BAFs in the interval that are greater than 0.5 and the mean of all BAFs in the interval that are less than 0.5. These values represent the mean BAF suggested by the data as reported using black bars in all BAF plots.

7.4 Tree Construction

We describe how the trees associated with the results from a run of THetA2 are constructed. First, this tree should not be interpreted as a phylogenetic tree, but rather as a tree representing the nested partitioning of inferred subpopulations and the aberrations whose population frequencies place them in each subpopulation. This construction of a binary tree partition is defined formally and studied in [3]. We only create such trees when they can be constructed unambiguously. This will always be the case for mixtures of three or fewer subpopulations, but since THetA2 makes no "perfect phylogeny" assumption about the subpopulations that it infers, such a tree may not be constructible with four or more subpopulations.

Each tree is constructed as follows. Each subpopulation is a leaf and is annotated with the fraction of the tumor mixture that was predicted to account for that population. For any pair of tumor subpopulations that share aberrations we add a parent node connecting them and label the node with the total fraction of cells in the sample that are part of either subpopulation. We iterate this process up the tree until we can join all remaining populations with a root node. The aberrations labeled on leaf nodes are unique to that subpopulation. Any aberrations that are shared among the tumor subpopulations are labeled on their parent node, rather than labeling each leaf node. An aberration is listed as a whole-arm event when more than a fixed proportion (> 0.7) of the chromosome arm was predicted to be either deleted or amplified in a single subpopulation. Finally the root of the tree represents the complete collection of cells in the sample.

8 TCGA Samples: Additional Results

Table S2 contains a complete list of genomes analyzed broken down by TCGA sample ID and the available datatypes and purity estimates for each. Table S3 contains the complete purity estimation results across all samples, including TCGA histopathology results and purity estimates reported for the ABSOLUTE algorithm [2].

8.1 Whole Exome Sequencing Data

When considering a tumor to be a mixture of normal cells and a single tumor population we find that THetA2 purity estimates obtained from both the ExomeCNV and EXCAVATOR interval segmentations to be similar for most genomes (Supplemental Fig. S8) with a few outliers. While the Pearson correlation coefficient between the purity estimates obtained from the different segmentations is 0.47, most of this error comes from two samples, TCGA-06-0185 and TCGA-AO-A0JF, and the correlation increases to 0.9 when these two samples are excluded. THetA2 infers multiple tumor subpopulations in sample TCGA-06-0185, so we surmise that the discrepancy between the purity estimates is due to the presence of subclonal copy number aberrations. We infer that sample TCGA-AO-A0JF contains copy number aberrations in a small subpopulation (Supplemental Fig. S9) by running THetA with parameters that allow for normal contamination up to 100% cells (rather than using the default settings). We believe this leads to the discrepancy in purity estimates between the two segmentation methods when run with the default parameters. We therefore exclude this sample from further analysis.

8.2 Consistency Across Sequencing Platforms

For the 7 genomes for which we have both whole-exome and whole-genome data, we compare THetA2 results across both data types. To compare copy number predictions, we use two different similarity metrics (see Supplementary Table S3). For similarity metric 1 (CNA Sim 1) we calculate the fraction genomic intervals in $\mathbf{I}^*$ where THetA2 returns the same copy number for the whole-genome and whole-exome data. For similarity metric 2 (CNA Sim 2) we relax the assumption that THetA2 returns the same

Table S2: **Genomes analyzed** - A list of the genomes analyzed, the cancer type and what type of datasets were available for sample purity analysis. ABS refers to ABSOLUTE results obtained from SNP array data as reported in [2].

ID	Cancer Type	ABS	WXS	WGS	WGS (low)
TCGA-06-0137	GBM	X	X		
TCGA-06-0145	GBM	X	X		
TCGA-06-0171	GBM	X	X		
TCGA-06-0174	GBM	X	X		
TCGA-06-0185	GBM	X	X	X	
TCGA-06-0188	GBM	X	X	X	
TCGA-06-0214	GBM	X	X	X	
TCGA-06-0219	GBM	X	X		
TCGA-06-2557	GBM		X		
TCGA-56-1622	LUSC		X	X	
TCGA-A2-A0EU	BRC		X		X
TCGA-AO-A0JF	BRC		X		X
TCGA-AO-A0JJ	BRC		X		X
TCGA-AO-A0JL	BRC				X
TCGA-BH-A0W5	BRC		X		X
TCGA-13-1500	OV	X	X		
TCGA-29-1768	OV	X	X		
TCGA-A3-3324	KIRC			X	

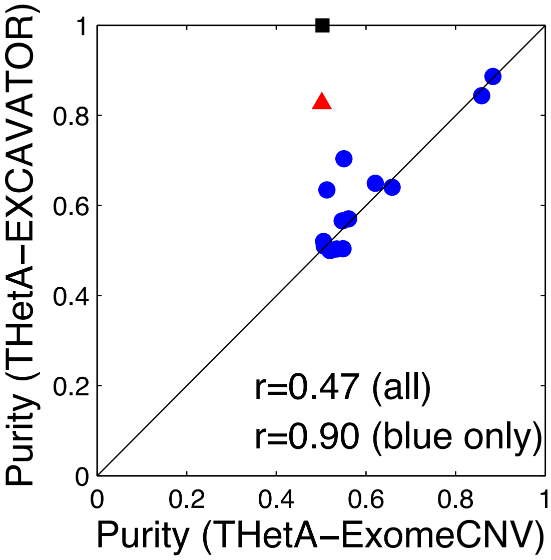


Figure S8: **Comparison of purity estimates obtained for two different whole-exome segmentation methods when considering a tumor to be a mixture of normal cells and one tumor population.** The sample indicated by the red triangle is TCGA-06-0185. The sample indicated by the black square is TCGA-AO-A0JF. Values of r shown are the Pearson correlation coefficient over either all the datapoints, or the indicated subset.

Table S3: **Comparison of THetA2 results on whole-genome and whole-exome data.** Path. are purity estimates reported in TCGA histopathology reports. ABS are ABSOLUTE purity estimates reported by [2]. WGS Purity, WXS Purity and # populations are values predicted by THetA.* indicates that the sample did not pass the criteria to be considered for multiple tumor populations (see Supplemental Material - Interval Selection). Overlap is $\frac{\mathbf{I}^*}{|\mathbf{I}_{WGS}| \cup |\mathbf{I}_{WXS}|}$ where $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$ are the interval partitions for the whole-genome and whole-exome data, respectively, and $\mathbf{I}^*$ is the set of intervals longer than 100kb contained in both $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$. CNA Sim1 is the fraction of $\mathbf{I}^*$ where the copy number estimates are the same between the two data types. CNA Sim2 is the fraction of $\mathbf{I}^*$ where the copy number estimates are of the same type (deletion, amplification, normal) between the two data types. [1]For sample TCGA-06-0214, WGS data was aligned to hg18 and WXS data aligned to hg19. We also compared to WGS data aligned to hg19, but found it contained a much larger variance in read depth than the hg18 data.

			WGS Purity	WXS Purity			
Sample	Path.	ABS	(# populations)	(# populations)	Overlap	CNA Sim1	CNA Sim2
TCGA-06-0137	0.85-0.9	0.92	-	0.89 (2*)	-	-	-
TCGA-06-0145	0.8-0.9	0.79	-	0.84 (2*)	-	-	-
TCGA-06-0171	0.3-0.5	0.76	-	0.68 (3)	-	-	-
TCGA-06-0174	0.8-0.9	0.95	-	0.92 (3)	-	-	-
TCGA-06-0185	0.95	0.89	0.87 (3)	0.83 (2*)	0.97	0.91	0.91
TCGA-06-0188	0.6-0.8	0	0.70 (3)	0.63 (3)	0.96	0.79, 0.62	0.80, 0.70
TCGA-06-0214[1]	0.25-0.8	0.66	0.67 (3)	0.67 (3)	0.96	0.97, 0.92	0.97, 0.94
TCGA-06-0219	0.8-0.95	0.65	-	0.69 (3)	-	-	-
TCGA-06-2557	1.0	-	-	0.58 (3)	-	-	-
TCGA-56-1622	0.9	-	0.68 (3)	0.78 (3)	0.96	0.89, 0.57	0.91, 0.77
TCGA-A2-A0EU	0.9	-	0.77 (3)	0.90 (3)	0.91	0.61, 0.22	0.64, 0.31
TCGA-AO-A0JF	0.7	-	0.52 (2*)	1.00 (2*)	0.97	0.98	0.98
TCGA-AO-A0JJ	0.8	-	0.52 (3)	0.52 (2)	0.85	0.67	0.68
TCGA-AO-A0JL	0.8	-	0.87 (3)	-	-	-	-
TCGA-BH-A0W5	0.7	-	0.51 (2*)	0.54 (2*)	0.98	0.97	0.97
TCGA-13-1500	0.89	0.75	-	0.77 (3)	-	-	-
TCGA-29-1768	0.25-0.5	0.55	-	0.87 (3)	-	-	-
TCGA-A3-3324	0.3-0.45	-	0.58 (2*)	-	-	-	-

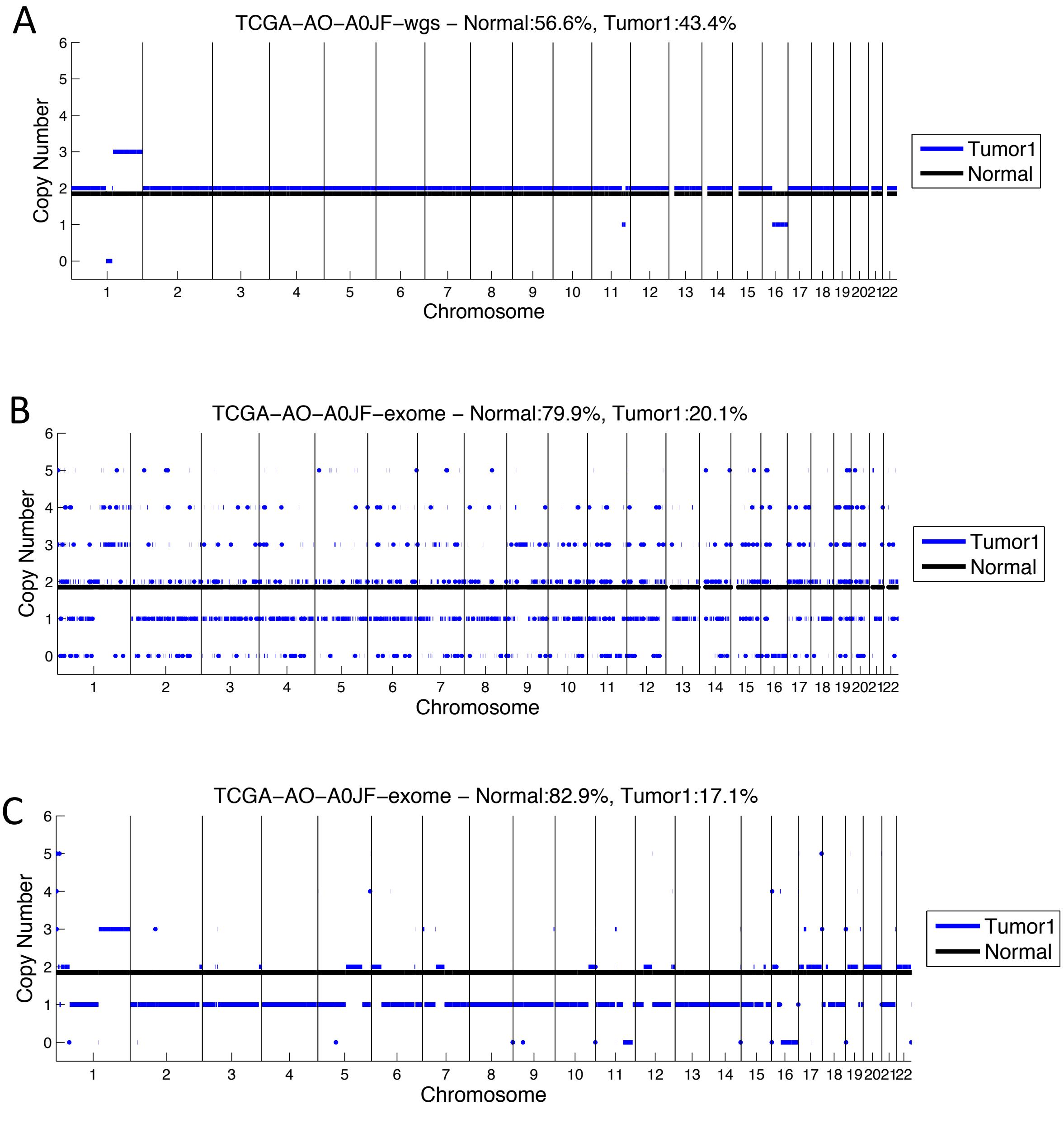


Figure S9: **THetA2 results when analyzing whole-genome and whole-exome data for sample TCGA-AO-A0JF and considering normal contamination up to 100% cells. A.** Results using the BIC-Seq partition from whole genome data. **B.** Results using the ExomeCNV partition on whole-exome data. **C.** Results using the EXCAVATOR partition on whole-exome data. All 3 indicate that sample purity is < 0.5.

integer copy number in the whole-genome and whole-exome data, and instead calculate the fraction of intervals in $\mathbf{I}^*$ where the copy state (normal, deleted, amplified) is the same for both datatypes. To account for different numbers of populations predicted from the different datatypes (either due to different estimates or one datatype not passing all criteria of multiple population analysis), we report similarity between the two largest subpopulations, and when applicable, the similarity between the two smaller subpopulations.

8.3 Sample TCGA-06-0188

We perform additional analysis on GBM sample TCGA-06-0188 which reported ABSOLUTE results [2] indicate as non-clonal and therefore was unable to determine sample purity. TCGA histopathology reports this sample as having purity between 0.6-0.8. Both whole-genome and whole-exome data from TCGA was available for this sample. THetA results on whole-genome data indicate that the sample contains 30% normal cells and two tumor populations in 43.2% cells and 26.8% cells (Supplemental Figure S10A). Results from applying THetA to whole-exome data are similar and indicate that the sample contains 36.6% normal cells and two tumor populations in 43.1% cells and 20.3% cells (Supplemental Figure S10B). Notably, both purity estimates are within the range indicated by histopathology. A number of large copy number aberrations are predicted from both data types. Virtual SNP array analysis appears to indicate the existence of aberrations predicted by both data types, such as clonal deletion of 13q and subclonal deletion of 10 as well as other aberrations inferred from the whole-exome data such as clonal amplification of chromosome 7, clonal deletion of chromosome 22q and subclonal deletion of 17p (Supplemental Figure S10C).

8.4 Low-Pass Breast Cancer Genomes

We include here the sample composition inferred by THetA2 for two of the low-pass breast cancer genomes, TCGA-A2-A0EU and TCGA-AO-A0JL (Supplemental Figure S11), for which we infer multiple distinct tumor subpopulations. Both genomes appear highly rearranged and we predict a number of chromosome arm events. We note that our inferred purity values of 0.77 and 0.88 are near the reported histopathology purity values of 0.9 and 0.8 for these samples.

8.5 Sample TCGA-56-1622

We present further results for squamous cell lung cancer sample TCGA-56-1622. First, Figure 5(b) in the main manuscript shows the observed read depth over 50kb bins as well as the predicted read depth as determined by the inferred tumor composition for this sample. For a given vector $\mathbf{x}$ we define $\widehat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|}$. Predicted read depth is calculated using $\mathbf{C}$, μ and normal read depth vector $\mathbf{w}$. Let W be the square matrix with entries along the main diagonal equal to $\mathbf{w}$ and all other entries 0. The predicted read depth ratio for interval j is: $\frac{\widehat{(W\mathbf{C}\mu)}_j}{\widehat{(\mathbf{w})}_j}$.

Second, using THetA2 results, we are able to identify several deletions and amplifications that have been reported as recurrent CNAs in lung cancers [4] (see Supplemental Table S4). In particular, we see a high amplification in 3q26 (see Supplemental Figure S12). Amplification in this region has been reported to be particularly common in squamous cell carcinoma genomes, and contains several genes which have been identified as potential oncogenic drivers in squamous cell carcinoma, including PI3KCA, SOX2, p63, SSCRO/DCUND1, and TERC [4].

8.6 Sample TCGA-06-0214

For this sample, we ran THetA2 with $n = 2, 3, 4$ on the whole-genome data. We find that after correcting for model size using the BIC, the $n = 2$ and the $n = 4$ solutions have a lower likelihood than the $n = 3$ solution.

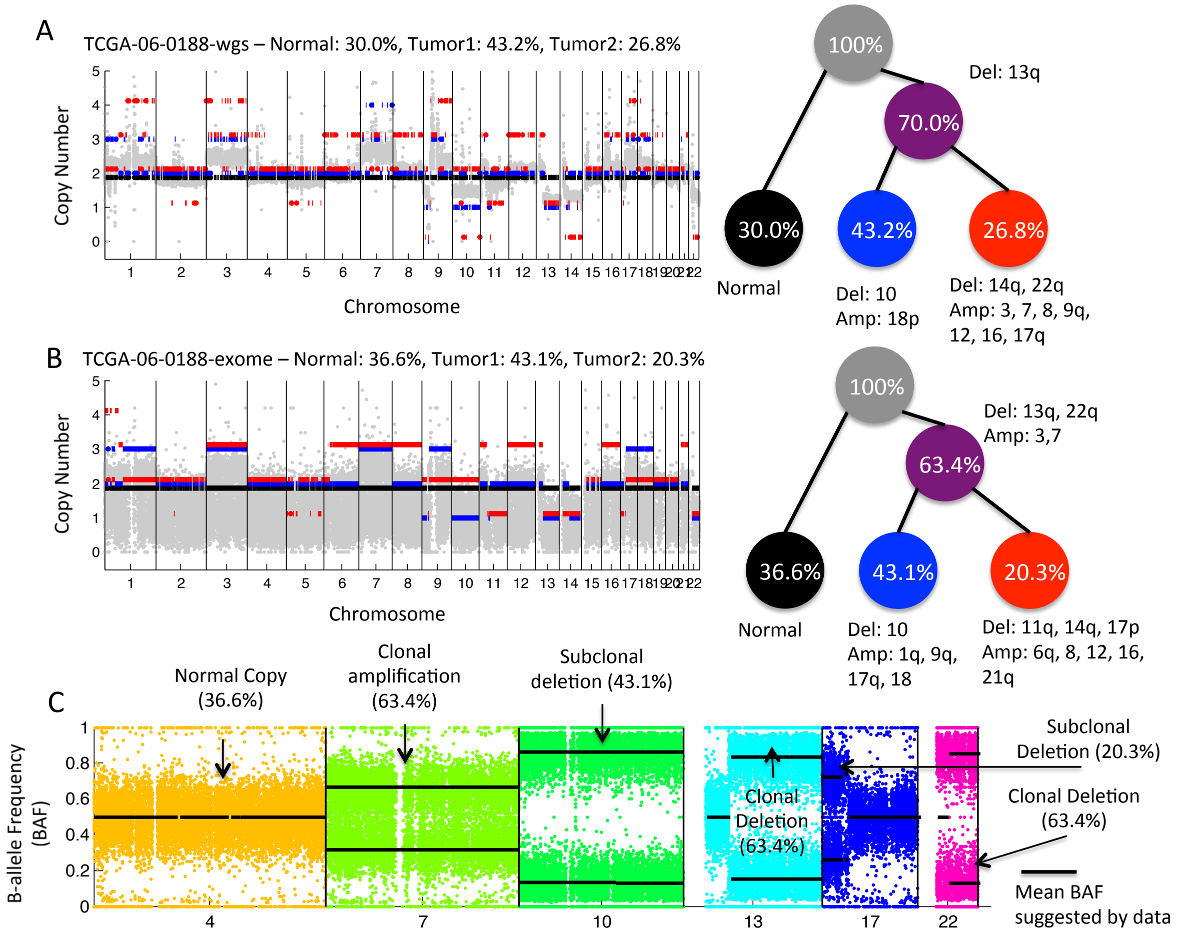


Figure S10: **THetA2 results when analyzing whole-genome and whole-exome data for sample TCGA-06-0188. A.** (Left) Read depth rations (gray) over 50 kb bins and the inferred copy number aberrations for intervals > 2 Mb calculated by THetA2 applied to whole-genome data when the tumor is considered to be a mixture of 3 subpopulation: normal cells (black), and two tumor subpopultions (blue and red). (Right) A reconstruction of the tumor mixture along with ancestral clonal population (purple) with the inferred aberrations and estimated fraction of cells in each population. THetA2 results when analyzing whole-genome data for sample TCGA-06-0188. **B.** Same as the previous part, but applied to whole-exome data. **C.** Virtual SNP array showing B-allele frequencies for chromosomes 4, 7, 10, 13, 17 and 22.

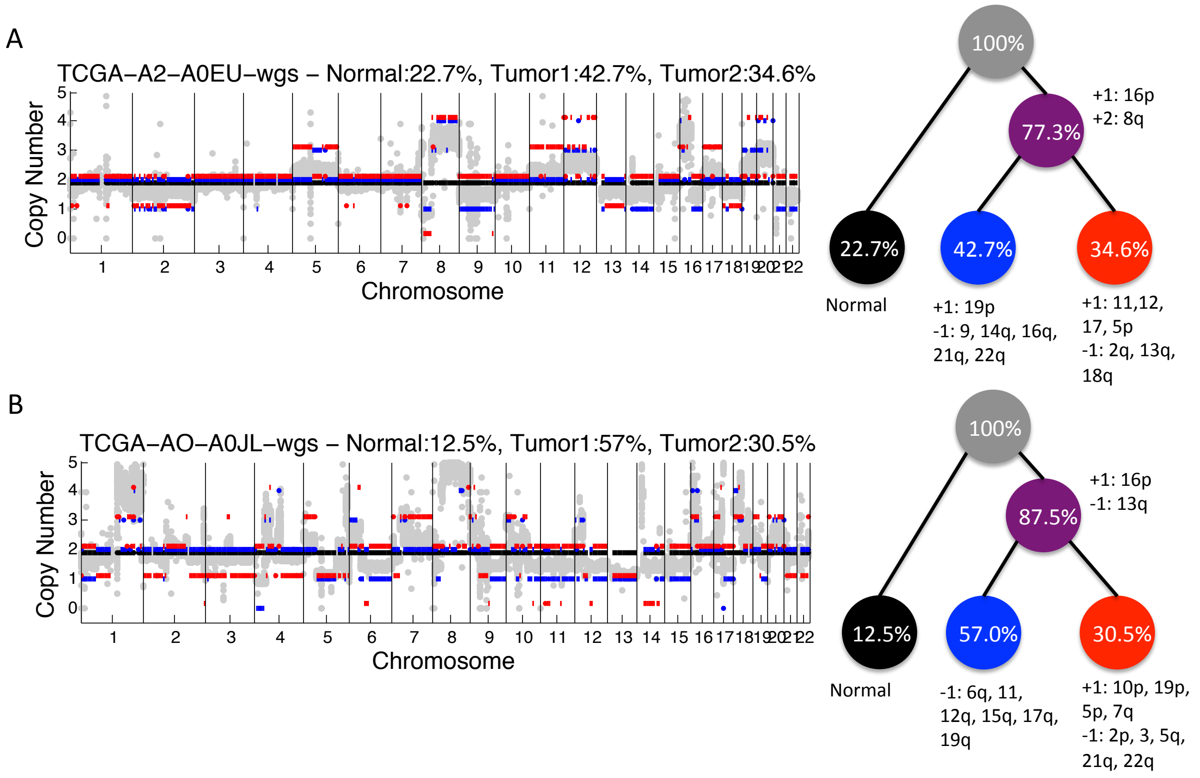


Figure S11: **THetA2 results when analyzing low pass whole-genome data for two breast cancer samples predicted to have 3 subpopulations from low pass whole-genome data.** Read depth ratios (gray) over 50 kB bins and the inferred copy numbers (for all intervals > 2Mb) for a mixture of normal cells (black) and two distinct tumor subpopulations (blue and red) inferred by THetA2.

ID	CNA Type	Tumor 1 (50%)	Tumor 2 (18.1%)
3q26	Amplification	X	X
5p13-14	Amplification		X
8q23	Amplification		X
8q24	Amplification	X	X
3p21	Deletion	X	X
8p21	Deletion	X	X
9p21-22	Deletion	X	X
13q22	Deletion	X	
17p12-13	Deletion	X	

Table S4: A list of clonal and subclonal CNAs identified in squamous cell lung cancer sample TCGA-56-1622 by THetA2 which have been reported as recurrent CNAs in lung cancers [4].

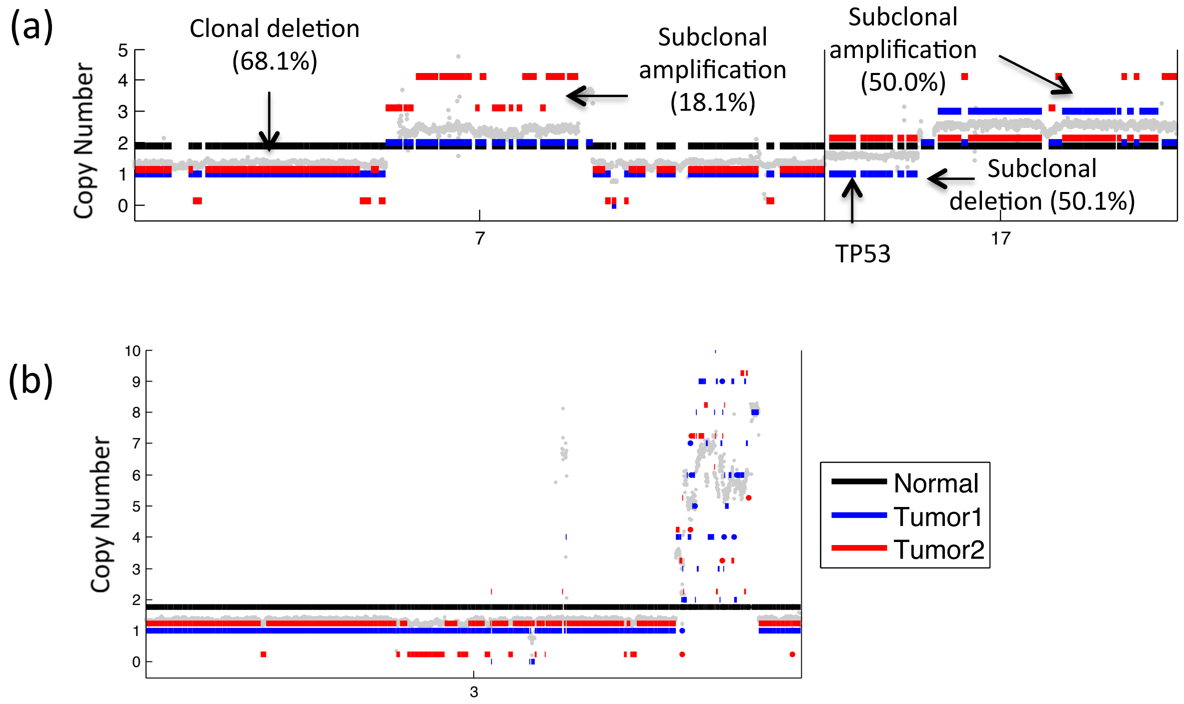


Figure S12: **THetA2 results when analyzing whole-genome for sample TCGA-56-1622.** **(a)** Zoomed in view of chromosomes 7 and 17 where we identify several copy number aberrations including a subclonal deletion containing TP53. **(b)** Zoomed in view of chromosome 3. We are able to identify several CNAs common in squamous cell lung cancer, including deletion in 3p21, and amplification in 3q26.

8.7 Sample TCGA-06-0145

For glioblastoma sample TCGA-06-0145, THetA outputs two possible $(\mathbf{C}, \mu)$ pairs using only read depth – one largely haploid and one largely diploid. We apply our probabilistic model of BAFs described previously and find that the diploid reconstruction, which includes rearrangements characteristic to glioblastoma such as amplification of chr7 and deletion of chr10 [10], is determined to be the more likely tumor composition (see Supplemental Fig S13).

References

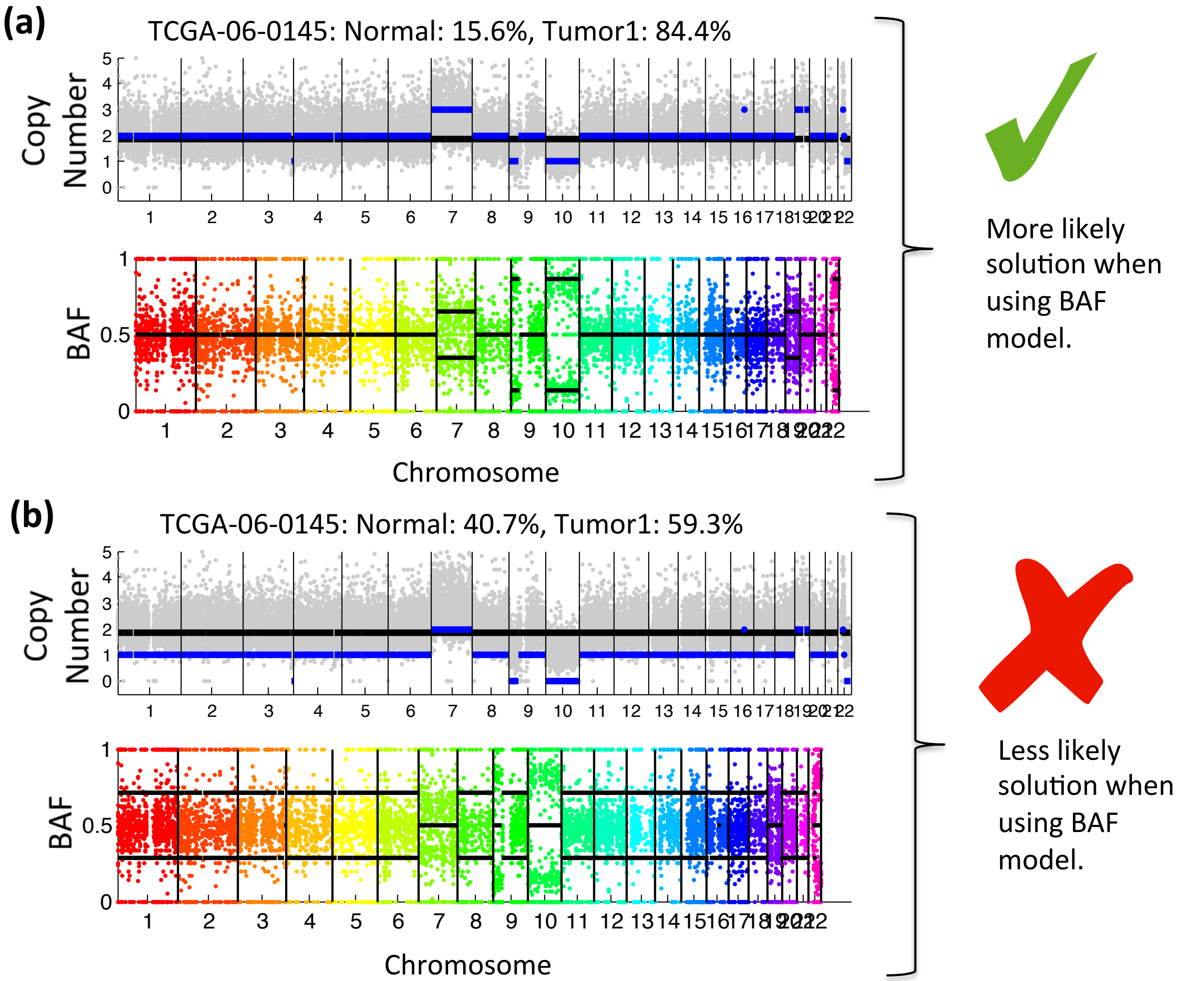
[1] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–74, May 2013.

[2] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic dna alterations in human cancer. *Nat Biotechnol*, 30(5):413–21, May 2012.

[3] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86, Jun 2014.

[4] Rebecca S Heist, Lecia V Sequist, and Jeffrey A Engelman. Genetic changes in squamous cell lung cancer: a review. *J Thorac Oncol*, 7(5):924–33, May 2012.

[5] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler,



TCGA-06-0145: Normal: 15.6%, Tumor1: 84.4%

More likely solution when using BAF model.

TCGA-06-0145: Normal: 40.7%, Tumor1: 59.3%

Less likely solution when using BAF model.

Figure S13: **Analysis of two equally likely solutions returned by THetA2 for GBM sample TCGA-06-0145. (a)** One reconstruction returned by THetA2. (Top) Read depth ratios over 50kb bins (gray) and inferred copy numbers for normal genome (black) and one cancer genome (blue). (Bottom) Observed BAF for the genome along with expected BAF calculated using $(\mathbf{C}, \mu)$. Under the BAF model described in Equation (2) this reconstruction is determined to be more likely. **(b)** Same as (a) but for the second solution returned by THetA2. Under the BAF model described in Equation (2) this reconstruction is determined to be less likely.

Rachel A Harte, Steve Heitner, Angie S Hinrichs, Katrina Learned, Brian T Lee, Chin H Li, Brian J Raney, Brooke Rhead, Kate R Rosenbloom, Cricket A Sloan, Matthew L Speir, Ann S Zweig, David Haussler, Robert M Kuhn, and W James Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Res*, 42(Database issue):D764–70, Jan 2014.

[6] Alberto Magi, Lorenzo Tattini, Ingrid Cifola, Romina D'Aurizio, Matteo Benelli, Eleonora Mangano, Cristina Battaglia, Elena Bonora, Ants Kurg, Marco Seri, Pamela Magini, Betti Giusti, Giovanni Romeo, Tommaso Pippucci, Gianluca De Bellis, Rosanna Abbate, and Gian Franco Gensini. Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biol*, 14(10):R120, Oct 2013.

[7] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol*, 14(7):R80, Jul 2013.

[8] Jarupon Fah Sathirapongsasuti, Hane Lee, Basil A J Horst, Georg Brunner, Alistair J Cochran, Scott Binder, John Quackenbush, and Stanley F Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv. *Bioinformatics*, 27(19):2648–54, Oct 2011.

[9] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222–30, Jun 2009.

[10] Dominik Sturm, Sebastian Bender, David T W Jones, Peter Lichter, Jacques Grill, Oren Becher, Cynthia Hawkins, Jacek Majewski, Chris Jones, Joseph F Costello, Antonio Iavarone, Kenneth Aldape, Cameron W Brennan, Nada Jabado, and Stefan M Pfister. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer*, 14(2):92–107, Feb 2014.

[11] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang, Mark D Johnson, Donna M Muzny, David A Wheeler, Richard A Gibbs, Raju Kucherlapati, and Peter J Park. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc Natl Acad Sci U S A*, 108(46):E1128–36, Nov 2011.